# Replicability of Classification Procedures for Gene Expression Data

---

Thesis
Presented in Partial Fulfillment of the Requirement for the Graduation of Research Distinction with a Degree in Electrical and Computer Engineering in the College of Engineering of The Ohio State University.

Dinank Gupta
The Ohio State University
2017

Advisor: Mohammadmahdi R. Yousefi, Department of Electrical and Computer Engineering

# ABSTRACT

Pattern classification is a branch of statistics and machine learning that uses labeled samples to predict information about unlabeled ones. A common application of this theory in medicine is to classify cancer patients into subtypes based on the patterns of their gene expression profiles. What determines the validity of the procedure is not whether one can find these patterns in observed data, but whether these patterns generalize to unobserved data from the same population. In this regard, the error of the classification rule over the population determines its validity and a key issue is how to estimate it.

In small sample situations, where the number of observed data is small, estimating the classification error becomes problematic as most of the error estimators have high variance. This raises doubts on the replicability of small-sample studies. In this thesis, I will use a replicability index to asses multiple classification and error estimation procedures that are commonly used in the medical community, and in particular, on RNA-seq and microarray gene expression data, and provide suggestions on the sample size to ensure that a procedure applied to a small preliminary study will generalize in a large follow-on study with an acceptable margin of error.

## ACKNOWLEDGEMENT

First of all, I would like to greatly thank my advisor, Dr. M. R. Yousefi. Thanks to his constant support and guidance, I was able learn a lot and make my undergraduate career a great experience. I am thankful to him for all the lessons that he gave me and for persistently helping me learn new concepts. Also, I would extend my gratuity towards him for helping me become a better research writer by helping me through various drafts of this thesis among many other writings during this time.

I would like to thank the Office of Undergraduate Research and Creative Enquiry for their constant support throughout my research journey. I specially thank Paige Trojanowski who helped me in the beginning of my research career.

Lastly, I thank my friends and family for listening to me talk about my work and supporting me throughout this time.

# TABLE OF CONTENTS

# LIST OF TABLES AND FIGURES

# INTRODUCTION

Many recent cancer diagnosis and prognosis studies have suggested that gene expression profiles, such as microarrays and RNA-seq data, may serve as reliable detectors/predictors of several cancers or their outcomes. While trying to discover these clinically useful biomarkers, a preliminary study with a small set of specimens is conducted, then a statistical analysis or machine learning procedure is carried out and if the results are satisfactory, a follow-on study with a large set of samples is conducted to validate the findings.

Notwithstanding the expectation, it has been estimated that as much as 75% of published results are not replicable [14], which implies that most of the reported biomarkers perform poorly on data other than the one they were designed on. There has been an increasing discussion about the reproducibility crisis [2,5,9] and the quality and generalizability of biomedical research [4].

Reproducibility issues are associated to the measurement platform, specimen handling, sample compatibility between studies and their normalization [5]. Replicability is concerned with the statistical analysis and inference methods chosen to either make decisions or support/rule out potential hypotheses. The former can be alleviated through the standardization of sample processing and data sharing protocols. Mitigation of the latter, however, requires a careful analysis of statistical significance and error. In the context of cancer diagnosis and prognosis, classifier error estimation and its accuracy is key to replicability of a medical test.

In my research, I analyze the replicability of classification procedures used on microarray and RNA-seq data. Microarray is a relatively old technology that measures the expression levels of multiple genes simultaneously, using image processing techniques, while RNA-seq is a technology that uses next-generation sequencing (NGS) to measure the abundance of RNA transcripts, and thus gene expression, in a biological sample. Given a collection of gene expression profiles from two or more different biological conditions, such as healthy individuals and cancer patients, statistical and machine learning methods can be used to classify a sample into one of the conditions. Replicability of an *interesting* preliminary medical classifier involves assessing its validity based on a large independent follow-on study.

It is often expected that not only the follow-on study will confirm findings of the preliminary study, but also provides a more accurate error estimate. However, it has been reported that most of the medical studies are not replicable, meaning that the reported error rate is significantly off [2,6, 17]. The statistical and machine learning communities have long known that the sample size affects the classification error rate, and more importantly the accuracy of its estimator. Collecting biological samples is expensive. This constrains the size of a typical preliminary study to be small, which jeopardizes the validity of the entire classification and error estimation procedure. It is thus crucial to provide a formal framework to address generalizability of a classification procedure, without wasting valuable resources on testing classifiers that are simply not good.

We use a probabilistic quantity, termed as the *replicability index*, to characterize how often a follow-on gene expression classification study yields results that are at least as good as those in the preliminary study. More importantly, if a researcher desires a certain level of replicability for a preliminary study, this quantity can be used to suggest the minimum sample size necessary to guarantee that level. We test this index on a multitude of classification procedures and data generation models that mimic real-world gene expression studies.

# METHODOLOGY

(Gene *Expression Data Models*)

Two models are used to generate synthetic expression data. Both the models are built using parameterized multivariate distributions, each representing a biological condition. Sample points are generated from two classes with $p$ features. Thus, each sample point is specified using a feature vector $\boldsymbol{X} \in R^p$ and a label $K \in \{0,1\}$. Let $n$ denote the number of patients (or, samples) in the preliminary small sample study. So, a study of 20 patients with 10,000 genes each means that number of samples are 20 and number of features are 10,000. number of features are 10,000.

### Microarray Model

For microarray simulation, we are using the model developed by Hua et al., 2005 [8]. The class-conditional densities are multivariate Gaussian with

$f(x|K=k) \sim N_p(\mu_k, \Sigma_k)$ for $k \in \{0,1\}$ where $\mu_0$ and $\Sigma_0$ are $p\times1$ mean vector and $p\times p$ covariance matrix of class 0, and $\mu_1$ and $\Sigma_1$ are the mean vector and covariance matrix of class 1. To set the values of these parameters, we break them into smaller groups of parameters that signify biological properties of microarray samples in cancer studies.

We assume features belong to one of the two groups of markers and non-markers. Markers are the genes associated with a disease or condition related to the disease and they have different class-conditional distributions for the two classes. They are further categorized into two subgroups: global markers and heterogeneous markers [7]. Global markers take values from $D_{gm}$- dimensional Gaussian distribution with parameters $(\mu_0^{\mathrm{gm}}, \Sigma_0^{\mathrm{gm}})$ for sample points from class 0 and $(\mu_1^{\mathrm{gm}}, \Sigma_1^{\mathrm{gm}})$ for sample points from class 1. Heterogeneous markers, on the other hand, are divided into two subgroups of equal size, each associated with one of two mutually exclusive subclasses within class 1. Therefore, a sample belonging to 1 of the subclass takes values from $D_{hm}$- dimensional Gaussian distribution with parameters$\left(\mu_1^{\mathrm{hm}}, \Sigma_1^{\mathrm{hm}}\right)$. The same markers for sample points belonging to the other subclass and class 0 take values from a Gaussian distribution with parameters $(\mu_0^{\mathrm{hm}}, \Sigma_0^{\mathrm{hm}})$.

We assume identical covariance structures for both global and heterogeneous markers, meaning that $\Sigma_0^{\mathrm{gm}} = \Sigma_1^{\mathrm{gm}} = \Sigma_0^{\mathrm{hm}} = \Sigma_0^{\mathrm{hm}} = 0.62\,\Sigma$. We assume that $\Sigma$ has a block matrix structure with zeros off the diagonal, and identical $5\times5$ matrices, $\Sigma_\rho$, on the diagonal. Further, $\Sigma_\rho$ itself has 1 on the

diagonal and 0.4 for the rest. As for the mean vectors, we assume global markers and the heterogeneous markers have the same mean vectors: $\{\mu_0^{gm}, \mu_1^{gm}\} = \{\mu_0^{gm}, \mu_1^{gm}\} = \{\mu_0, \mu_1\}$. Furthermore, we set $\mu_0 = [0,0,\dots,0]^T$ and $\mu_1 = [\theta,\theta,\dots,\theta]^T$. Here $\theta$ may be fixed or may be random, in which case it has a probability distribution.

Non-markers are also divided into two subgroups: high-variance non-markers and low-variance non-markers [7]. High-variance non-markers are uncorrelated and they follow the distribution given by a Gaussian mixture distribution $v\ N(0,0.62) + (1-v)N(\theta,0.62)$, where $v \sim Unif(0,1)$. Low-variance non-markers are also uncorrelated with identical one dimensional Gaussian distribution with $N(0,0.62)$. Figure 1 represents the block-based structure of the model. Simulation parameters for this data model are also listed in Table 1.

| **Parameters** | **Value** |
|---|---|
| Feature Size | 20,000 |
| Training sample size (equal size classes) | 40 to 200 |
| Testing sample size | 5000 |
| Mean of class 1 $(\theta)$ | Fixed: 1. 167<br>Random: $N(1.167,0.036)$ |
| # Global Markers | 20 |
| # Subclasses | 2 |
| # Heterogeneous Markers per Subclass | 50 |
| # High-variance non-markers | 2000 |
| # Low-variance non-markers | 17880 |

**Table 1**: Parameters for microarray simulations

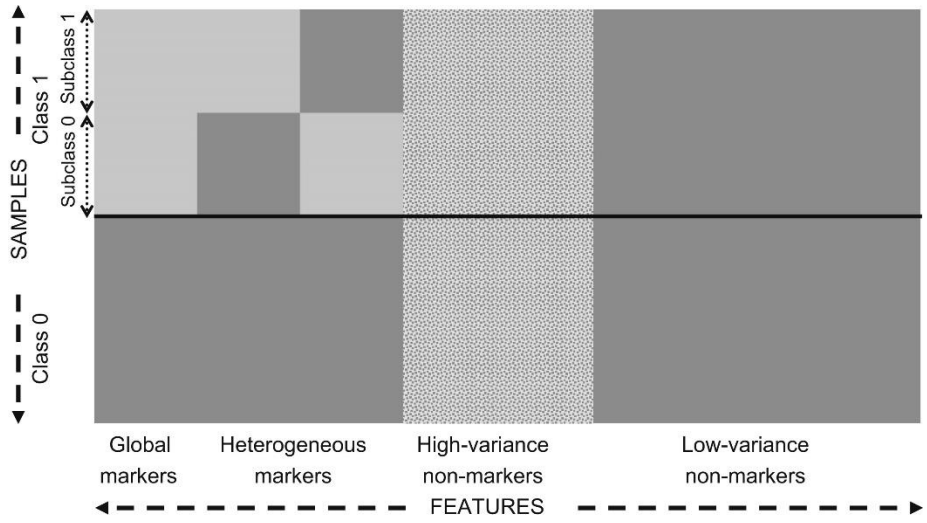


**Figure 1**: Block matrix structure for microarray data

## RNA-seq Model

In this work, we use RNA-seq model developed by Witten, *et al.,* 2012 [16]. We denote RNA-seq data in the form of an $n \times p$ matrix, $\boldsymbol{X}$, with each element representing the number of reads in a gene. $\boldsymbol{X}$ represents $n$ observations with $p$ features. This data is usually assumed to follow a Poisson log linear [12] or Negative Binomial distributed model [1].

Just like microarray, here we assume that observations are drawn from population containing two biological conditions. The model can be written as:

$$f(x|K = k) \sim NB\left(N_{ij}, \phi_j\right),$$

where $N_{ij} = s_i g_j d_{kj}$ [16] is the mean of the distribution and $\phi_j$ is the dispersion parameter. This factor $s_i$, (also called *size factor/sequencing depth)* will be used to reflect that different samples may be sequenced to different depths. [12,13] The factor $g_j$ reflects the variability in the total number of reads per sample. $d_{kj}$ (also called *differentially expressed parameter)* is used to reflect if features are differentially expressed (DE) [3,12]. The probability of a feature being DE is 0.01

or 0.05, depending on the simulation. If a feature is differentially expressed, then $\log(d_{kj})$ is Gaussian with a given mean and variance. Also, for features that are DE, we introduce a parameter $\theta$ to control the mean of their distribution.

| Parameters | Value |
|---|---|
| Feature Size | 10,000 |
| Training sample size (equal size classes) | 40 to 200 |
| Testing sample size | 5000 |
| $s_i$ | $Unif(0.2, 2.2)$ |
| $g_j$ | $Exp(1/25)$ |
| $d_{kj}$ for all k if not differentially expressed | 1 |
| Mean of $\log(d_{kj})$ if differentially expressed $(\theta)$ | Fixed: 0<br>Random: $N(0,0.036)$ |
| Variance of $\log(d_{kj})$ if differentially expressed | 0.0025 |
| Dispersion Parameter $(\phi)$ | 0.1 |

**Table 2**: Parameters for RNA-seq simulations

In real situations, sequencing data tends to be over-dispersed compared to a Negative Binomial model. This can be accounted for by transforming the data using a power transformation with parameter $\alpha \in (0,1]$ [11,16]. To this end, the transformation $X'_{ij} \leftarrow X_{ij}^{\alpha}$ is done, where $\alpha$ is chosen such that:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(X'_{ij} - X'_{i\cdot}\,X'_{\cdot j}/X'_{\cdot\cdot}\right)^2}{X'_{i\cdot}\,X'_{\cdot j}/X'_{\cdot\cdot}} \approx (n-1)(p-1),$$

where $X'_{i\cdot} = \sum_{j=1}^{p} X_{ij}$ , $X'_{\cdot j} = \sum_{i=1}^{n} X_{ij}$ and $X'_{\cdot\cdot} = \sum_{i,j} X_{ij}$.

(Classification *Techniques*)

Classifier rule model is defined as a pair $(\Psi, \Xi)$, [17] where $\Psi$ is a classification rule, and $\Xi$ is an error estimation rule on training data of the feature-label distribution $F$. In a typical classification task, a random training set $S_n = \{(X_1, K_1), (X_2, K_2), \ldots, (X_n, K_n)\}$ [7] is drawn from $F$ and then a classifier $\psi_n = \Psi(S_n)$ is designed which takes $\boldsymbol{X}$ as input and outputs a label $K$. The true

classification error is given as $\varepsilon_n = P(\psi_n(\boldsymbol{X}) \neq K)$. [7]. Then an error estimation rule Ξ is used to estimate the error as $\hat{\varepsilon}_n = \Xi(S_n)$ for a classifier $\psi_n$.

### 1. Linear Discriminant Analysis (LDA)

LDA is a plug-in rule for Bayes classifier when class densities are Gaussian with a common covariance matrix. LDA assigns the label 1 to sample point $\boldsymbol{X}$, if and only if

$$\left(\boldsymbol{X} - \overline{\mu}_1\right)^T \hat{\Sigma}^{-1} \left(\boldsymbol{X} - \overline{\mu}_1\right) \leq \left(\boldsymbol{X} - \overline{\mu}_0\right)^T \hat{\Sigma}^{-1} \left(\boldsymbol{X} - \overline{\mu}_0\right),$$

where $\overline{\mu}_k$ is the sample mean for classes $k \in \{0,1\}$, and $\hat{\Sigma}$ is the sample covariance matrix. Since LDA is a classifier that is designed on a Gaussian distributed data, we expect it to perform well on microarray gene expression data.

### 2. Sparse-Poisson Linear Discriminant Analysis (Sparse-PLDA)

If we assume that the data is coming from a Poisson or a Negative Binomial distribution, we expect a model based on a similar distribution to perform better. This is a common assumption made for RNA-seq data [11,15,16]. We apply s-PLDA exclusively to RNA-Seq data. We estimate the parameter given in the method in the following manner: We estimate $g_j$ as $\hat{g}_j = \boldsymbol{X}_{\cdot j}$. The size factors $s_1, s_2, \ldots s_n$ are estimated using Maximum Likelihood Estimator (MLE) for $N_{ij}$ as $\hat{s}_i = \boldsymbol{X}_{i\cdot}/\boldsymbol{X}_{\cdot\cdot}$. We also use a shrinkage estimator for $\hat{d}_{kj}$ as:

$$\hat{d}_{kj} = \begin{cases} \frac{a}{b} - \frac{\rho}{b} & if\ \sqrt{b}\left(\frac{a}{b} - 1\right) < \rho \\ \frac{a}{b} + \frac{\rho}{b} & if\ \sqrt{b}\left(1 - \frac{a}{b}\right) > \rho \\ 1 & if\ \sqrt{b}\left|1 - \frac{a}{b}\right| < \rho \end{cases}$$

where $C_k \in \{0,1\}, a = X_{C_k} + \beta$ , $b = \sum_{i \in C_k} \widehat{N}_{ij} + \beta$, $X_{C_k} = \sum_{i \in C_k} \boldsymbol{X}_{ij}$ and $\widehat{N}_{ij} = \hat{s}_i \hat{g}_j$. Here $\rho$ is a non-negative tuning parameter that is found using cross validation. In the simulations, we assume $\beta = 1$.

Let $y^*$ be the unknown label of testing sample. The classifier finds the probability of the testing sample $X_j^*$ having a label $k$ by [16]:

$$\log \mathrm{P}\left(y^* = k \middle| X_j^*\right) = \sum_{j=1}^{p} X_j^* \log \hat{d}_{kj} - \hat{s}^* \sum_{j=1}^{p} \hat{g}_j \hat{d}_{kj},$$

where $\hat{s}^*$is the sequencing depth estimate of the testing sample. We assign the label $k$ to the $y^*$ value with higher probability. Here, to simplify the analysis, we assume that both the classes have equal prior probability.

## 3. *k*-Nearest Neighbor

*k*-NN classifies a new sample using *k* nearest training data samples nearest to the testing sample. Both the nearest 'distance' and the number of samples *k* can be adjusted while classifying the data.

## 4. Support Vector Machines

SVM finds a maximal margin hyperplane for a given set of training sample points [7]. If data cannot be separated by a linear function, some slack variables can be introduced in the optimization process to solve the problem. Otherwise,

one can transform the data by projecting it into a higher-dimensional space, where it is linearly separable.

(*Error Estimators*)

### *1.* $k$ Fold - Cross Validation

In $k$F-CV, the training sample, $S_n$ is randomly partitioned into $k$ folds $S_n^i$ for $i = 1,2, \dots k$ [7]. Now, each fold is held out the classifier design and then used as a test set. A classifier $\psi_n^i$ is designed on the remaining sets $S_n \backslash S_n^i$ and the error of the classifier is estimated by counting the misclassified sample points. The error is estimated by averaging the error for each fold. To reduce the variance due to random selection of the partitions, we repeat this process 10 times and then take the mean of the error as the $k$F-CV estimated error. We take $k = 5$ for our simulations.

### *2.* Resubstitution

In resubstitution error estimator we train and test the classifier on the same training data to find the error in classification. This estimation technique is very heavily biased on the training samples and thus we expect poorest estimates using this method.

### *3.* 0.632 Bootstrap

In bootstrap error estimator, a bootstrap sample is made by drawing $n$ equally likely points with replacement from the original data. Then a classifier is designed

on the bootstrap sample and its error is calculated by finding the number of misclassified samples. Then a bootstrap zero estimate is found by finding the expected value of this error with respect to the bootstrap sampling distribution. We find this using 100 independent bootstrap samples. In 0.632 bootstrap estimator, we also add the estimate from resubstitution to incorporate for biasing of bootstrap towards training samples. Thus,

$$0.632\ \hat{\varepsilon}_{boot} = (0.368 * \hat{\varepsilon}_{resub}) + (0.632 * \hat{\varepsilon}_{zero}).$$

(*Replicability Index*)

We use a probabilistic measure for whether a classifier that shows promising results in a small sample study will perform well on a larger independent sample [17]. We say that the original study will replicate with accuracy $\rho \geq 0$ if $\varepsilon_n \leq \hat{\varepsilon}_n + \rho$.

Any error estimate will not lead to a follow-on study, since the estimated error should be sufficiently small to motivate further analysis of the classification procedure. We will use $\tau$ as a threshold value of error, such that the follow-up study occurs if and only if $\hat{\varepsilon}_n \leq \tau$. We define *replicability index* as:

$$R_n(\rho, \tau) = P(\varepsilon_n \leq \hat{\varepsilon}_n + \rho | \hat{\varepsilon}_n \leq \tau).$$

The pipeline to estimate the *replicability index* is as follows:

1. Generate the training and testing data.
2. Design classification method on the small size training samples.
3. Find the true error and estimated error of the designed classifier.

4. Find the *replicability index* based on required values of threshold and accuracy parameters.

The pipeline is also shown in Figure 2.

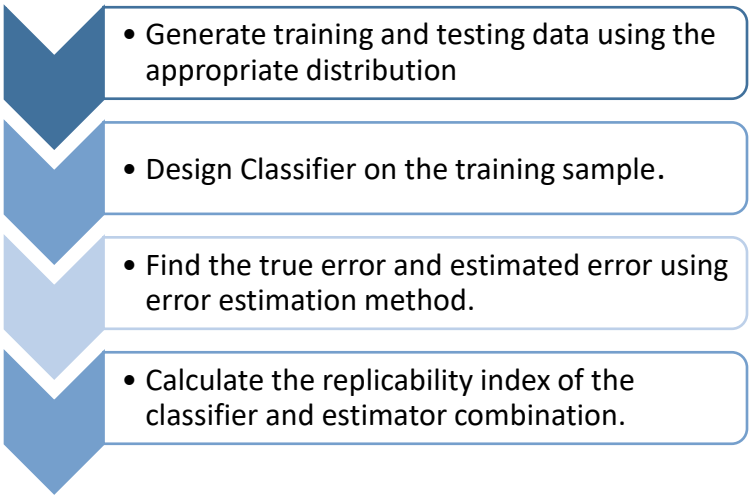


**Figure 2**: Pipeline to find the *replicability index*

## RESULTS AND DISCUSSION

Based on the steps mentioned in the previous section, we determined the r*eplicability index* of multiple classifier – error estimator combinations. We use LDA, SVM and 5-NN on simulated microarray data and s-PLDA on the simulated RNA-seq data. We apply cross validation, 0.632-bootstrap and resubstitution to find the estimated error of the combination.

(Microarray *Simulations*)

Figure 3 shows the expected *replicability index*, when $\theta$ is random, with respect to different sample size $n$, $\rho$ and $\tau$. We observe that there is a direct

correlation between expected *replicability index* and the training sample size. But we also observe that after a certain sample size, we don't see much improvement in the value of *replicability index*. Therefore, this gives a researcher an estimate of number of sample that the study requires to get a desired expected *replicability* of their method. For example, if required replicability is 0.8 and $\tau = 0.1$ , $\rho = 0.03$, the sample size should be about 120.

Figure 4 shows the effect of $\tau$ and $\rho$ on the replicability of the classification procedure for $n = 40$ when $\theta$ is fixed. We see that as we increase the value of threshold parameter $\tau$, the estimated replicability index increases, which is expected. These plots give us an idea about the performance of our method and what threshold error values should we expect from it. This also gives a nice comparison amongst different classification and error estimation techniques and thus might help a researcher understand the replicability degree of their classification procedure.

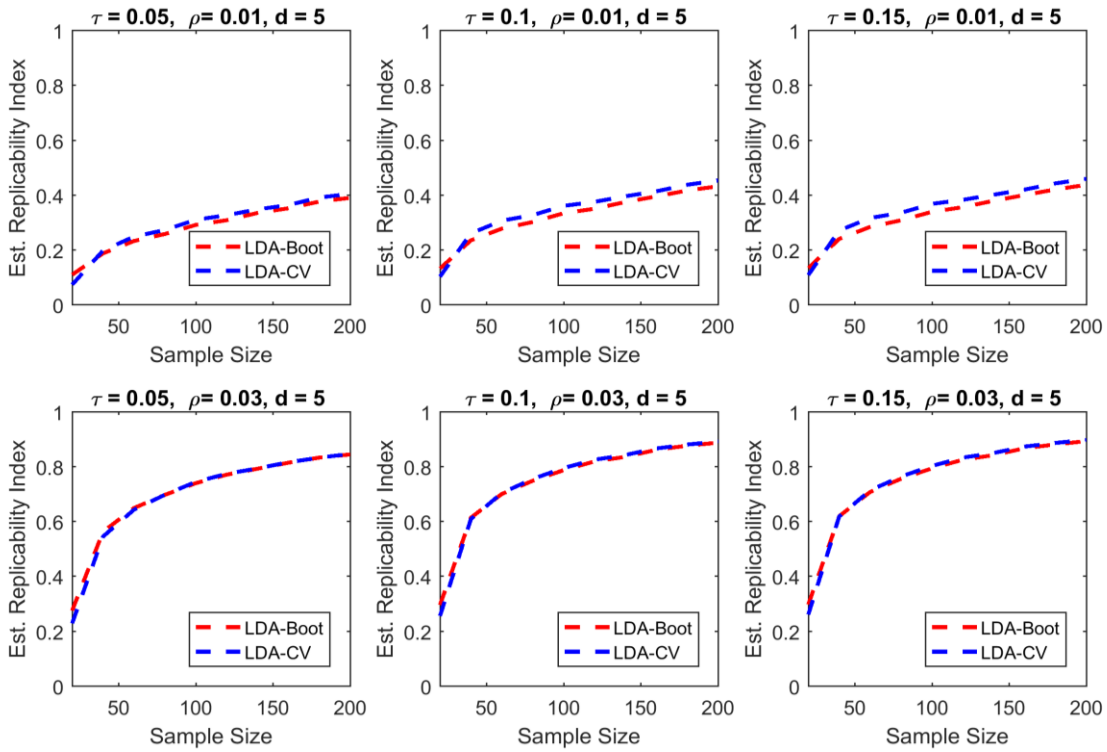


**Figure 3**: Replicability for LDA classification rule and 0.632 bootstrap and 5F-CV for microarray data ($d = 5$ features are selected using *t*-test): $\tau = 0.05, 0.1, 0.15$, $\rho = 0.01, 0.03$

Figure 5 has similar type of plots as Figure 4, but with $n = 120$ samples. We observe very similar trends in Figure 5 and Figure 4. The differentiating factor between the plots are the value of threshold parameter $\tau$ corresponding to a *replicability* value. In general, for $n = 120$, we observe that we get higher values of *replicability* for same value of $\tau$ than for $n = 40$, regardless of the analysis method. These results support the correlation between sample size and *replicability index,* we saw from Figure 3.

(*RNA-Seq Simulations*)

Figure 6 and 7 show replicability index plots for RNA-seq data. Figure 6 shows similar plots to Figure 3, when $\theta$ is random. We plot *replicability index* with respect to different sample size $n$, $\rho$ and $\tau$. Like Figure 3, *replicability* of classification procedure shows direct correlation to training sample size. But we observe that *replicability* in general is quite low, even for higher training sample size. Thus, we expect that even higher training size would give us better *replicability* values.

Figure 7 supports the observations from Figure 6. When we plot replicability index for various values of $\tau$ and $\rho$, for a fixed $\theta$, we see that for low $\tau$ values, replicability is near 0, meaning that our classifier error rates are pretty high in comparison to those from microarray. Also, we see that the higher probabily of DE features in the classifier, the better the classifier performance is. This is expected and can be attributed to more features being used in classifier training.

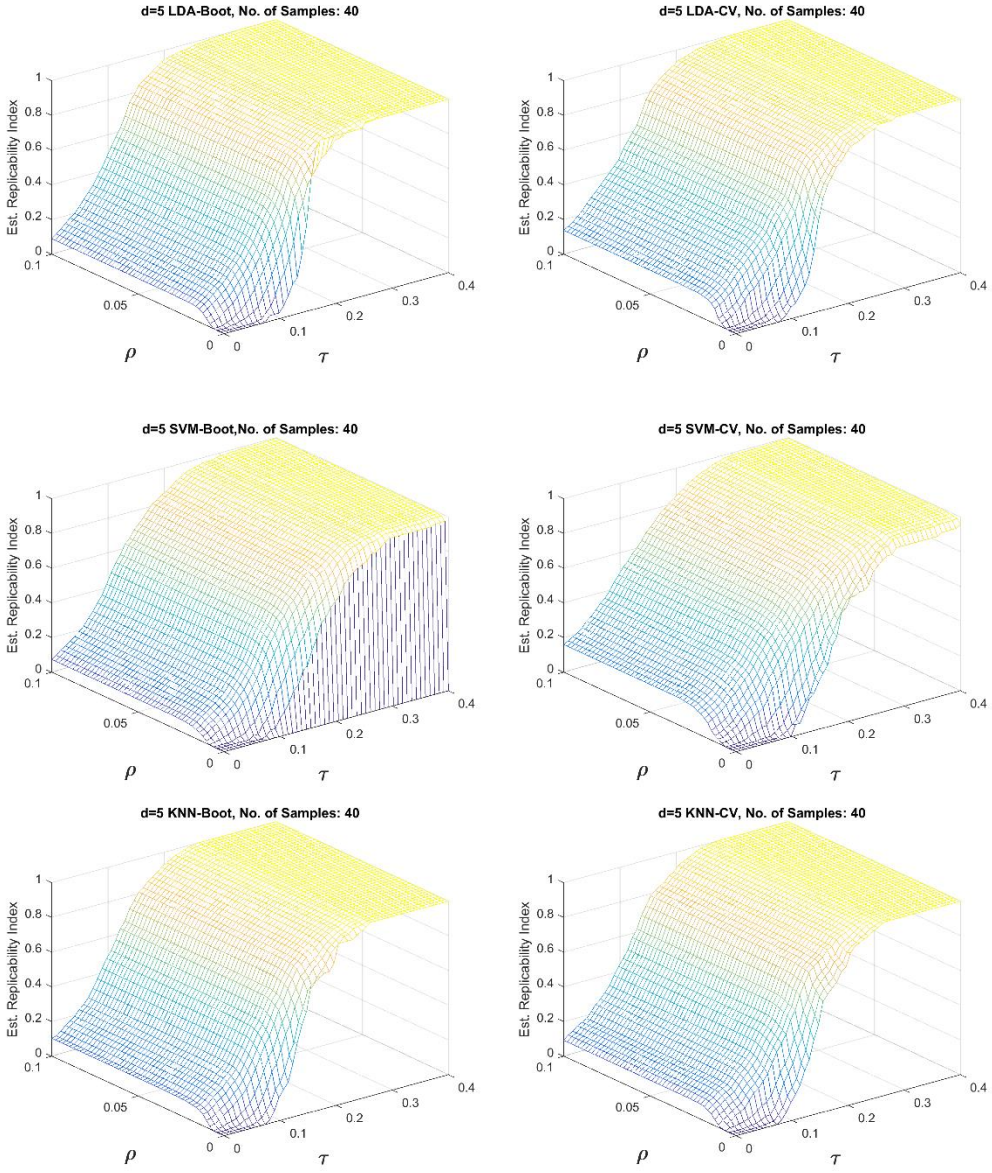


**Figure 4**: Replicability index for microarray using $n = 40$ using LDA, SVM, 5NN classification rule, 0.632 bootstrap, 5F-CV estimator rule for multiple values of $\tau$ and $\rho$. $d = 5$ features are selected using *t*-test

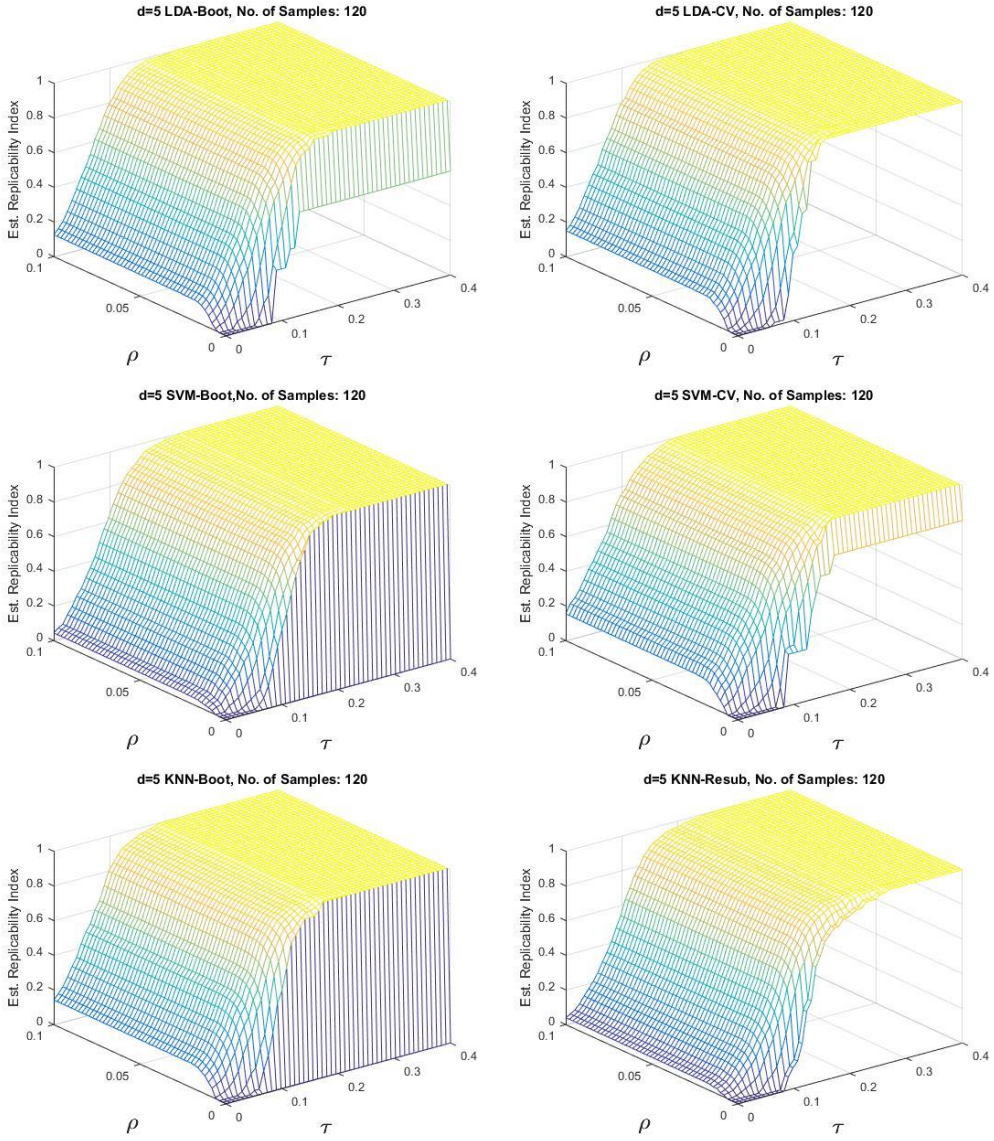


**Figure 5**: Replicability index for microarray using $n = 120$ using LDA, SVM, 5NN classification rule, 0.632 bootstrap, 5F-CV estimator rule for multiple values of $\tau$ and $\rho$. $d = 5$ features are selected using *t*-test

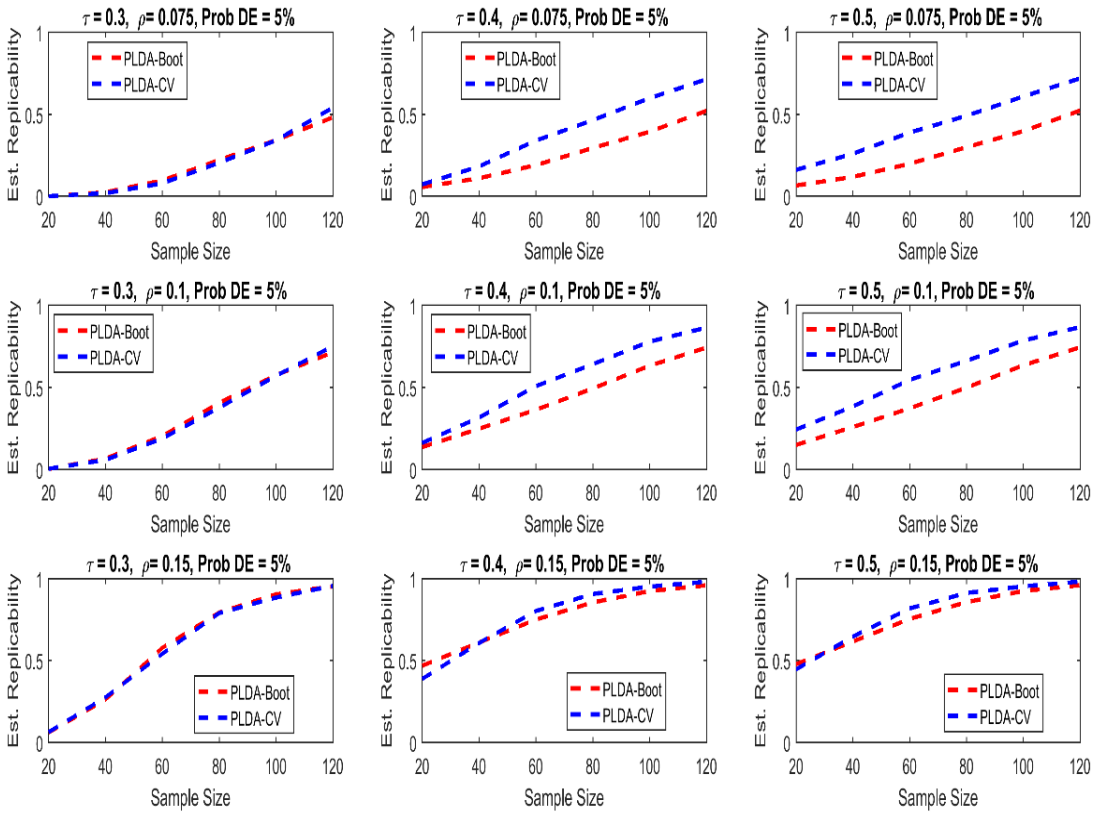


**Figure 6**: Replicability for RNA-seq: s-PLDA classification rule and 0.632 bootstrap and 5F-CV with $\tau = 0.3, 0.4, 0.5$, $\rho = 0.075, 0.1, 0.15$

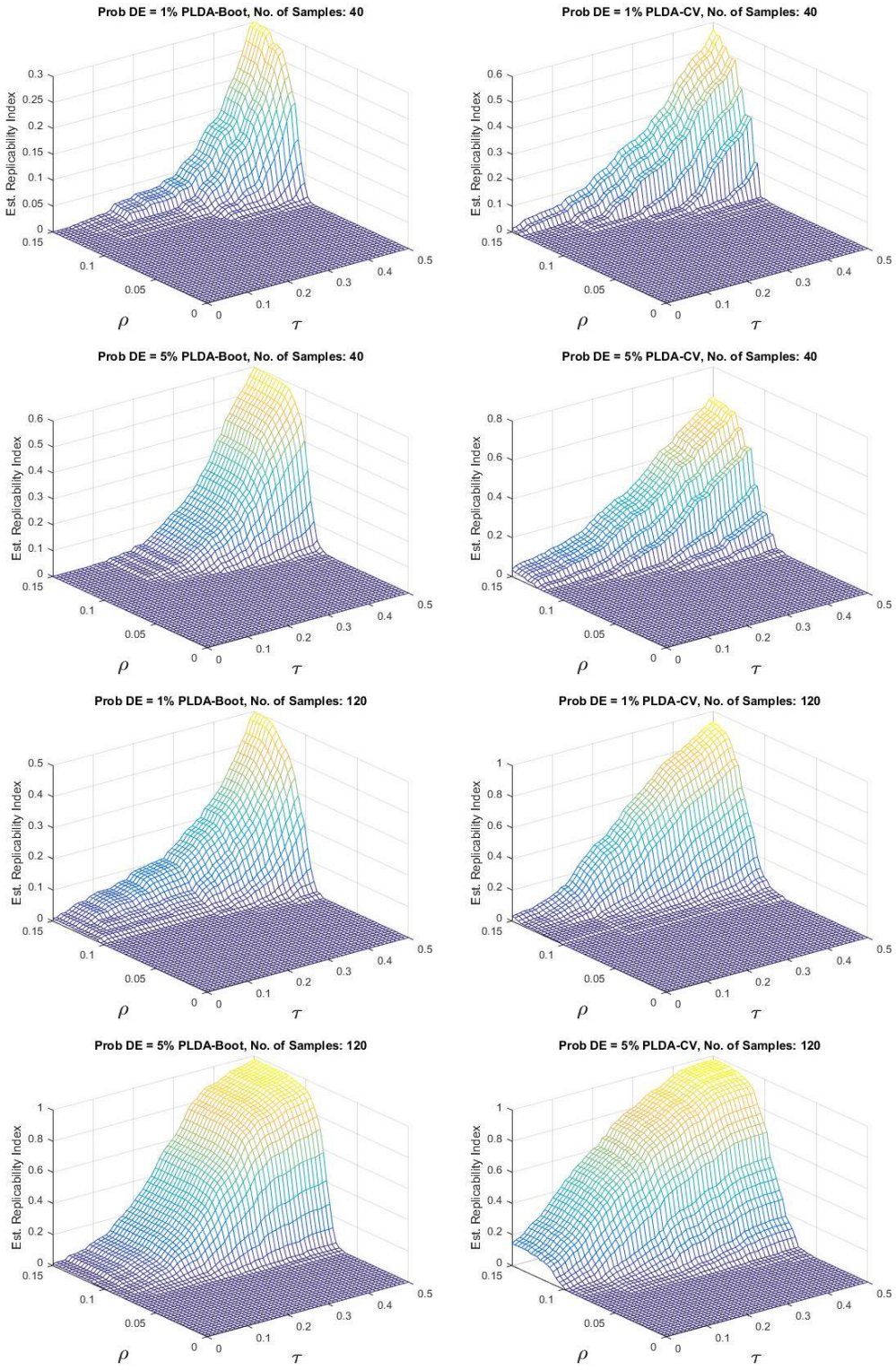


**Figure 7**: Replicability index for RNA-seq using $n = 40, 120$ using s-PLDA classification rule, 0.632 bootstrap, 5F-CV estimator rule for multiple values of $\tau$ and $\rho$.

# CONCLUSION

Performance replicability is an epistemological issue for any classification study. Replicability addresses the issue of accuracy of an error estimate. Larger the difference between estimated and true error, further we are from knowledge about the conclusion from preliminary study. If there is no measure of replicability, we have no justification to support a follow-up study. Thus, the issue of replicability should be settled prior to any large sample study.

In this project, we studied about the issue in replicability of classification in gene expression data. The issue of replicability is an important topic to discuss in this case because the problem of small sample is one that cannot be overcome, due to simply the cost of such experiments and the time required to collect medical samples. For this kind of study, we assume some prior knowledge about the data. It may be argued that prior assumption may be erroneous and end results might have a different conclusion. But if sufficient knowledge is not accepted for an experiment, we are not ready to do the experiment.

If we assume some knowledge about the data and find replicability of our method, we might save much more resources by not being over optimistic about our method. This study provided information to a researcher conducting a small sample study about number of sample points required in a preliminary study to get an accurate estimate of replicability of the method. This would help get a better understanding and avoid potential over-optimism in accessing a research methods' reliability.

# REFERENCES


[1] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, p. R106, 2010.

[2] A.-L. Boulesteix, “Over-optimism in bioinformatics research,” *Bioinformatics*, vol. 26, no. 3, pp. 437–439, Feb. 2010.

[3] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments,” *BMC Bioinformatics*, vol. 11, p. 94, 2010.

[4] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, “The Illusion of Distribution-Free Small-Sample Classification in Genomics,” *Curr Genomics*, vol. 12, no. 5, pp. 333–341, Aug. 2011.

[5] E. R. Dougherty, “Biomarker development: Prudence, risk, and reproducibility,” *Bioessays*, vol. 34, no. 4, pp. 277–279, Apr. 2012.

[6] E. R. Dougherty, “Bootstrap Error Estimation for Classification with - Dougherty_CB_2010_preprint.pdf.” [Online]. Available: http://www.ece.tamu.edu/~ulisses/public/Dougherty_CB_2010_preprint.pdf. [Accessed: 05-Mar-2017].

[7] N. Ghaffari, M. R. Yousefi, C. D. Johnson, I. Ivanov, and E. R. Dougherty, “Modeling the next generation sequencing sample processing pipeline for the purposes of classification,” *BMC Bioinformatics*, vol. 14, p. 307, 2013.

[8] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, Apr. 2005.

[9] J. P. A. Ioannidis, “Why Most Published Research Findings Are False,” *PLOS Medicine*, vol. 2, no. 8, p. e124, Aug. 2005.

[10] J. M. Knight, I. Ivanov, and E. R. Dougherty, “MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification,” *BMC Bioinformatics*, vol. 15, p. 401, 2014.

[11] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani, “Normalization, testing, and false discovery rate estimation for RNA-sequencing data,” *Biostatistics*, vol. 13, no. 3, pp. 523–538, Jul. 2012.

[12] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.

[13] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat Meth*, vol. 5, no. 7, pp. 621–628, Jul. 2008.

[14] T. Ray, “FDA’s Woodcock says personalized drug development entering ‘long slog’ phase”, *Pharmacogen. Rep.,* Oct. 2011.

[15] D. M. Witten and R. Tibshirani, “Penalized classification using Fisher’s linear discriminant,” *J R Stat Soc Series B Stat Methodol*, vol. 73, no. 5, pp. 753–772, Nov. 2011.

[16] D. M. Witten, “Classification and clustering of sequencing data using a Poisson model,” *Ann. Appl. Stat.*, vol. 5, no. 4, pp. 2493–2518, Dec. 2011.

[17] M. R. Yousefi and E. R. Dougherty, “Performance Reproducibility Index for Classification,” *Bioinformatics*, p. bts509, Sep. 2012.